J Braz Comput Soc (2013) 19:411–422
DOI 10.1007/s13173-013-0121-y

ORIGINAL PAPER

Microalgae classification using semi-supervised and active learning based on Gaussian mixture models

Paulo Drews-Jr. · Rafael G. Colares · Pablo Machado · Matheus de Faria · Amália Detoni · Virgínia Tavano

Received: 12 November 2012 / Accepted: 16 August 2013 / Published online: 8 September 2013

Abstract Microalgae are unicellular organisms that have different shapes, sizes and structures. Classifying these microalgae manually can be an expensive task, because thousands of microalgae can be found in even a small sample of water. This paper presents an approach for an automatic/semi-automatic classification of microalgae based on semi-supervised and active learning algorithms, using Gaussian mixture models. The results show that the approach has an excellent cost-benefit relation, classifying more than 90 % of microalgae in a well distributed way, overcoming the supervised algorithm SVM.

Keywords Active learning · Semi-supervised learning · Microalgae classification

1 Introduction

Microalgae are unicellular organism that can be found in a variety of sizes, structures and forms. These characteristics allows us to classify microalgae into different phytoplankton taxonomic groups. Microalgae classification is relevant to biology and oceanology, because the description of microalgae species at a certain time and place is important to the understanding of how energy is transferred from the food chain base to higher trophic levels [5]. Furthermore, it reflects changes in fish stocks and the carbon cycle of a given environment. The classification of microalgae and characterization of the predominant taxonomic groups has a diversity of applications, such as understanding of a phytoplankton community's structure. A recent census of marine life [4] gathered research from more than 80 nations, and lasted one decade, in order to obtain a global benthic biomass map predicted to the seafloor, phytoplankton included.

Microalgae are classified in groups based on different characteristics, with huge morphological variations such as round, oval, cylindrical, and fusiform cells, as well as projections like thorns, cilia, etc. In addition to the taxonomic classification, phytoplankton organisms can be classified according to their sizes: picoplankton (0, 2–2 μm), nanoplankton (2–20 μm), and microplankton (20–200 μm). Specific composition, size structure and biomass studies about phytoplankton communities are being developed through the classic method of optic microscopy [13], in which an observer has to manually manipulate a small water sample requiring more than a day for a complete analysis.

The use of particle analyzers has been an important tool to obtain information about the aquatic environment. It intends to efficiently obtain data about density, composition and morphometry of phytoplanktonic organisms. Typically, this automatic equipment is composed of an optical system capable of distinguishing microalgae from other particles in the sample and capturing images, along with software that assists in the classification and visualization of the cells. An automatic, or even semi-automatic, approach to classifying microalgae would greatly benefit research on this topic. This work presents an approach to an automatic/semi-automatic classification of microalgae based on machine learning algorithms.

P. Drews-Jr. (✉) · P. Machado · M. de Faria
Centro de Ciências Computacionais, Universidade Federal do Rio Grande (FURG), Rio Grande, RS, Brazil
e-mail: paulodrews@furg.br

R. G. Colares
Departamento de Ciência da Computação, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, MG, Brazil
e-mail: rcolares@dcc.ufmg.br

A. Detoni · V. Tavano
Instituto de Oceanografia, Universidade Federal do Rio Grande (FURG), Rio Grande, RS, Brazil

The proposed approach combines two types of learning: semi-supervised and active. The first assumes that only a small part of data has known ranking a priori, and tries to use information from non-ranked data to improve the classification. The second, active, searches the non-ranked data for the one that provides the most information gain, and then asks the user the rank of that data. In this work, both learning types were combined to improve microalgae classification. The process is initialized with semi-supervised learning, and then is improved using active learning.

In order to acquire the microalgae data, a FlowCAM particle analyzer [15] was used. It is capable of obtaining information concerning microorganism in water samples. Four experts analyzed and ranked the obtained data in order to validate the proposed approach.

2 Related work

Most studies found in the literature try to classify plankton, which, although not exactly the focus of this work, shares some similarities with our goal. Blaschko et al. [1] presented a comparison of supervised approaches to learning and classifying plankton. Those approaches are used to classify larger organisms than the targets of this work, thus presenting a greater number of relevant features, facilitating the learning process. Furthermore, those approaches used extensive supervised data, which makes it very costly and not extensible. Finally, Blaschko et al. [1] also used the FlowCAM and the best results obtained were around 70 %.

Another work of interest was proposed by Xu et al. [21], which uses a restrict set of supervised data classified with a SVM classifier, using non-ranked data to improve the learning. Although the presented approach is adequate to this work, it does not use experts as an information source. They obtain the information through a density method technique, which is sensitivity to the microalgae size. Due to the small size of the microalgae used in this study, the amount of information is reduced, which makes this approach unfeasible.

The work of Sosik and Olson [19] used the FlowCytobot equipment to extract features from the phytoplankton, on a similar process to the FlowCAM. The results obtained in the automatic classification were around 68 and 99 %, depending on the type of organism that were classified. They obtained least significant results to smaller plankton, the focus of this work.

Another work from Hu and Davis [14] uses *co-occurrence matrices* techniques and SVM to classify plankton. Using both supervised learning techniques, they obtained around 72 % of accuracy.

The problem of classification of microalgae was addressed in the work of Drews, Jr., et al. [9], where Gaussian mixture models were used together with semi-supervised and active learning. The present work is an extension of the previous work, where the methodology is detailed. Furthermore, we present and discuss a more thorough experimental data acquired using FlowCAM.

3 Methodology

As explained on Sect. 1, this work uses an approach based on the combination of two learning types: semi-supervised and active, with the objective to classify microalgae. The first step of the work was to obtain the data of the microalgae using the FlowCAM. Given a water sample, this equipment is capable of finding and analyzing microalgae in order to identify up to 26 different features to compose the databases used in this work. This work used only seven of these features: ESD diameter, ESD volume, width, length, aspect ratio, transparency, and CH1 peak.

We selected the best of these features using the approach proposed by Peng et al. [16]; the method is an optimal first-order approximation of the mutual information criteria. The selected features are in accordance with FlowCAM software manual [11], which defines these seven features as good features in general cases. Four experts analyzed and classified these datasets in order to generate a ground truth to validate the proposed approach. Fig. 1 shows the FlowCAM interface.

The first step of this proposed method is the development of a semi-supervised algorithm to classify the microalgae. IN this step, the algorithm receives as input just a small sample of ranked data, wherein at least one instance of each class needs to be provided. This allow that the algorithm is able

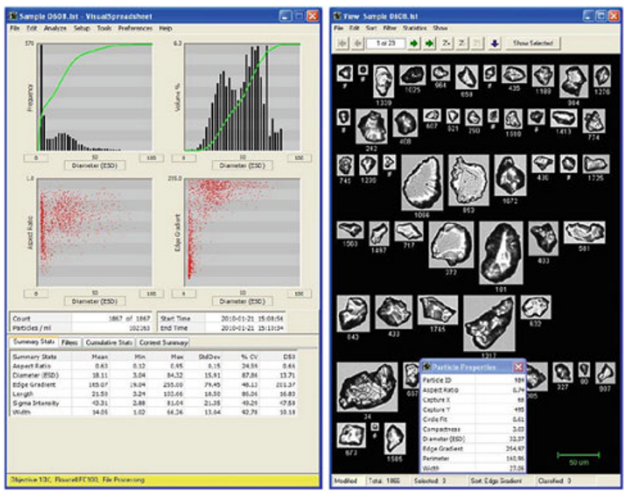

Fig. 1 FlowCAM interface [11]—The interface is divided into two windows. In the *left*, the Visual Spreadsheet is shown, where *tables, graphics* and *histograms* illustrate some statistics about the dataset. On the right, the *View Sample window* shows the microalgae images. The classification mechanism provided by FlowCAM is too simple and restricted to selecting limit values to features

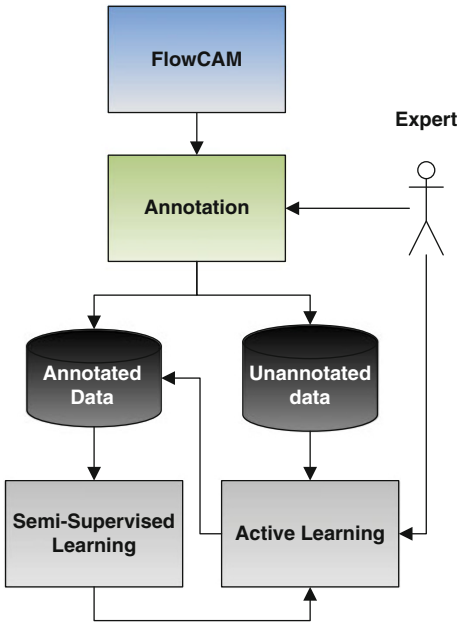

Fig. 2 Proposed approach

to identify and cluster microalgae with similar characteristics, creating a model of the microalgae class. This model allows new instances, non-ranked, to be observed and classified through their characteristics, updating the model simultaneously.

When the semi-supervised algorithm finishes, the active algorithm analyzes the instances that were not ranked and searches among them for the one that provides the largest information gain for the model. In order to identify which instance this is, three methods were used: *least confident sampling*, *margin sampling* and *entropy-based sampling*. Then, the chosen instance is presented to the user, who will indicate the class to which it belongs. This class is incorporated into the model, which is then updated and tries to classify the other non-ranked instances. This process is repeated as long as the user finds it to be favorable or until the information gain is too small. Figure 2 illustrates the described process. In the following sections, we describe the semi-supervised and active learning algorithms.

3.1 Semi-supervised learning

Due to the nature of the data used on this work, where the instances have similar characteristics when they belong to the same class, it is costly to rank a large set of instances. Thus, it favors an approach that uses clustering to classify microalgae. Furthermore, as the number of classes, species of microalgae on a sample are known and small,[1] and the classes are relatively well separated, the use of the *Gaussian mixture model* (GMM) with *expectation-maximization*(EM) becomes a natural choice [7].

3.1.1 Gaussian Mixture Models

The Gaussian mixture model (GMM) is a probability density function (PDF) given by a linear combination of a Gaussian PDF. More specifically, the function is a mixture of a Gaussian PDF if it has the following form:

$$p(x|K,\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \mathbf{\Sigma}_k), \tag{1}$$

where K is the number of the Gaussian PDF and π_k is the weight of each one in the mixture. This weight can be interpreted as the a priori probability that the random variable value was generated by the Gaussian k.

Considering, $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$, the GMM can be defined by the parameter list θ, which represents the parameters from each Gaussian and their respective weights, i.e., $\theta = \{\pi_1, \mu_1, \mathbf{\Sigma}_1, \ldots, \pi_K, \mu_K, \mathbf{\Sigma}_K\}$, where μ and $\mathbf{\Sigma}$ are the mean and the covariance matrix, respectively.

The problem with estimating the Gaussian mixtures lies in determining θ, given that only K and the data are known and the other parameters are unknown (π_k and $\theta_k = (\mu_k, \mathbf{\Sigma}_k)$).

Considering $Y = \{y_1, \ldots, y_n, \ldots, y_N\}$ with $y_n \in \mathbb{R}^M$, the independent sample set, where M is the size of the data sample space and N is the number of samples. In this work, y_n represents the dataset instances, the microalgae. It is possible to estimate the probability $p(y_n|K,\theta)$ directly for each K. However, a logarithmic function of the probability is normally used for ease of handling numbers. Thus, we have:

$$\hat{\theta} = \underset{\theta,K}{\operatorname{argmax}} \log p(y|K,\theta). \tag{2}$$

Solving the Eq. 2 is not an easy task [8,10]. The number of variables to be estimated can grow exponentially with the size of K and θ, thus making the computation very costly. We used the EM algorithm to solve this problem.

3.1.2 EM algorithm

The EM algorithm is used to determine the class of each data [7]. The algorithm aims to solve problems in which we do not know all the information needed for the solution.

The algorithm is composed of two steps:

E-step: On this step, the missing data are estimated using the observed data and the actual status of the model parameters.

M-step: The maximum likelihood function is maximized, considering that the missing data are known.

[1] This size is dependent of the environment. Typically, we have around ten different classes.

The EM algorithm seeks to classify the y_n data on classes, or Gaussian, and, later, to re-estimate each class value. Using Bayes' rule the probability that a point y_n belongs to class k is computed. Considering $\theta^{(i)}$ to be the θ value on the iterative step i of the algorithm and known in this step, the probability of *E-step* is given by:

$$p(k|y_n, \theta^{(i)}) = \frac{\pi_k \cdot \mathcal{N}(y_n|\theta_k^{(i)})}{\sum_{l=1}^{K} \pi_l \cdot \mathcal{N}(y_n|\theta_l^{(i)})}. \quad (3)$$

Calculating these probabilities makes it possible to estimate θ and π. These equations below show how each value is estimated in the maximization step (*M-step*). First, one normalizing parameter $\overline{N}_k$ is estimated by the posterior estimation of new values for $\overline{\pi}_k, \overline{\mu}_k, \overline{\mathbf{\Sigma}_k}$. From this, the update equations from step M can be defined:

$$\overline{N}_k = \sum_{n=1}^{N} p(k|y_n, \theta^{(i)}), \quad (4)$$

$$\overline{\pi}_k = \frac{\overline{N}_k}{N}, \quad (5)$$

$$\overline{\mu}_k = \frac{1}{\overline{N}_k} \sum_{n=1}^{N} y_n p(k|y_n, \theta^{(i)}), \quad (6)$$

$$\overline{\mathbf{\Sigma}}_k = \frac{1}{\overline{N}_k} \sum_{n=1}^{N} (y_n - \overline{\mu}_k) \cdot (y_n - \overline{\mu}_k)^{\mathrm{T}} \, p(k|y_n, \theta^{(i)}). \quad (7)$$

The algorithm initialization is critical for a good performance, i.e., the $\theta^{(0)}$. In this work, the initialization is done based on the ranked data available, generating a initial model using random initialization. Thereafter, using non-ranked data information, this model is updated. This approach has the major advantage of ensuring that data labeled as distinct classes remain this way.

The approach of Zhu and Goldberg [22] was used to estimate the GMM model from ranked and non-ranked data. The ranked data are computed distinctly in the E-step. This way, the ranked data have their probability set to 100 % for their class and 0 % to the other classes.

3.2 Active learning

After executing the semi-supervised learning algorithm, it is possible to divide the dataset into two groups: ranked instances and non-ranked instances. Considering $X = \{x_1, \ldots, x_n, \ldots, x_N\}$ as the set of non-ranked instances and k the possible classes, the active algorithm must find an $x_i \in X$ that maximizes the amount information added to the system when it is classified as k_j.

In order to define which instance x_i is going to be presented to the user, three metrics were used to calculate the information contained therein. The three metrics, based on the work of Settles [18] and Friedman et al. [12], are described below:

1. *Least-confident sampling*: involves choosing the instance with the least probability of belonging to the class with the most probability. The instance x to be chosen is the one that:

$$x = \underset{i}{\operatorname{argmin}} \; p(z_i = \hat{k}|x_i) \quad (8)$$

   where $\hat{k} = \operatorname{argmax}_k \; p(z_i = k|x_i)$ is the class with most probability.
2. *Margin sampling*: involves choosing the instance with the least margin between the class with most probability and the one with the secondmost probability. The instance x to be chosen is the one that:

$$x = \underset{i}{\operatorname{argmin}} \; [p(z_i = \hat{k}_1|x_i) - p(z_i = \hat{k}_2|x_i)] \quad (9)$$

   where $\hat{k}_1$ and $\hat{k}_2$ are the most likely classes.
3. *Entropy-based sampling*: involves choosing the instance with the most entropy of the classes' probabilities. The instance x to be chosen is the one that:

$$x = \underset{i}{\operatorname{argmax}} - \sum_{k} p(z_i = k|x_i) \log \; p(z_i = k|x_i) \quad (10)$$

After defining which instance is the most informative, the user must inform the system its rank. This classification is used by the EM algorithm in order to find the best model for the data, ranked or non-ranked, with this new information. Such model is initialized with the best representation until the present moment.

4 Experimental results

The results were obtained using two different datasets acquired using the FlowCAM equipment. The Oceanographic Institute of FURG collected the data during an oceanographic expedition on the Atlantic Ocean in different place and depth. In order to validate the results, four different experts classified these datasets. Doubtful data were eliminated, typically they were small microalgae, around 1 μm, or really big microalgae, which were problems on the acquisition by the FlowCAM or were microalgae colonies. Figure 3 illustrates some excluded data during the process.

The first dataset was classified on four different classes: flagellates (Fig. 4a), mesopores (Fig. 4d), pennate diatoms (Fig. 4c) and others (Fig. 4b). An important characteristic, usually found in this kind of data, is the unbalance of classes. The flagellates and the others classes represent more than 90 % of the data. Furthermore, as shown at Fig. 4a or b, these

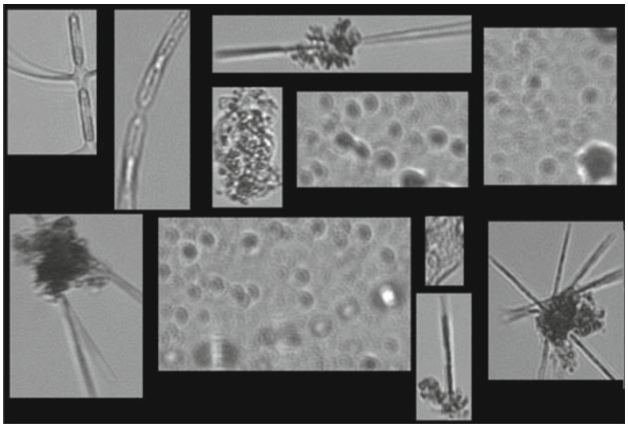

Fig. 3 Some examples of microalgae acquired by FlowCAM that were excluded due to acquisition problems or the presence of microalgae colonies. The presence of colonies is due to a failure in the segmentation process of FlowCAM. These fail are common due to the acquisition process of the FlowCAM device

are reduced size data with few characteristics, which makes the classification problem difficult to solve.

The second dataset was classified on four different classes: pennate diatoms (Fig. 5a), flagellates (Fig. 5b), gymnodinium (Fig. 5c) and prorocentrales (Fig. 5d), respectively. Both datasets have two similar species of microalgae and two different ones. This is due to different place and, mainly, depth where the samples were acquired. The characteristics of the data are similar, both datasets are unbalanced and with reduced size data.

In order to validate the proposed approach, we used some evaluation metrics. As there are multiple classes, the metrics need to deal with this kind of information. It was used the F-score metric [17], defined by the Eq. 11, which is the harmonic mean between the recall (r) and the precision (p), defined by Eq. 12.

$$F_\beta = (1+\beta)\frac{pr}{(\beta^2 p)+r}, \tag{11}$$

where β is a constant factor. At the present work, β was equal to 1, obtaining the F1-score metric.

$$r_k = \frac{Tp_k}{Tp_k + Fp_k}, \quad p_k = \frac{Tp_k}{Tp_k + Fn_k}, \tag{12}$$

where, Tp_k is the number of correctly classified microalgae for class k; Fp_k is the number of false positives, the number of microalgae that were wrongly classified as class k; Fn_k is the number of false negatives, the number of microalgae that are from class k, but were defined to another class; k is the microalgae class. These metrics are defined for each class.

The F1-score values are defined on the interval (0, 1), and if they are near one they represent a better classification, while small values, near zero, represent a low classification quality. However, to evaluate the performance for all classes was used the *micro-average* and *macro-average* metrics [20]. These metrics evaluate the average performance of the classifier, based on precision and recall metrics. The *macro-average* metric gives an average where every class is treated with same importance, while the *micro-average* metric gives an average where the microalgae are treated with the same importance.

It is important to evaluate these two metrics due the fact that the *micro-average* is more influenced by the classifier performance on classes with large samples, while the *macro-average* is more influenced by classes with less

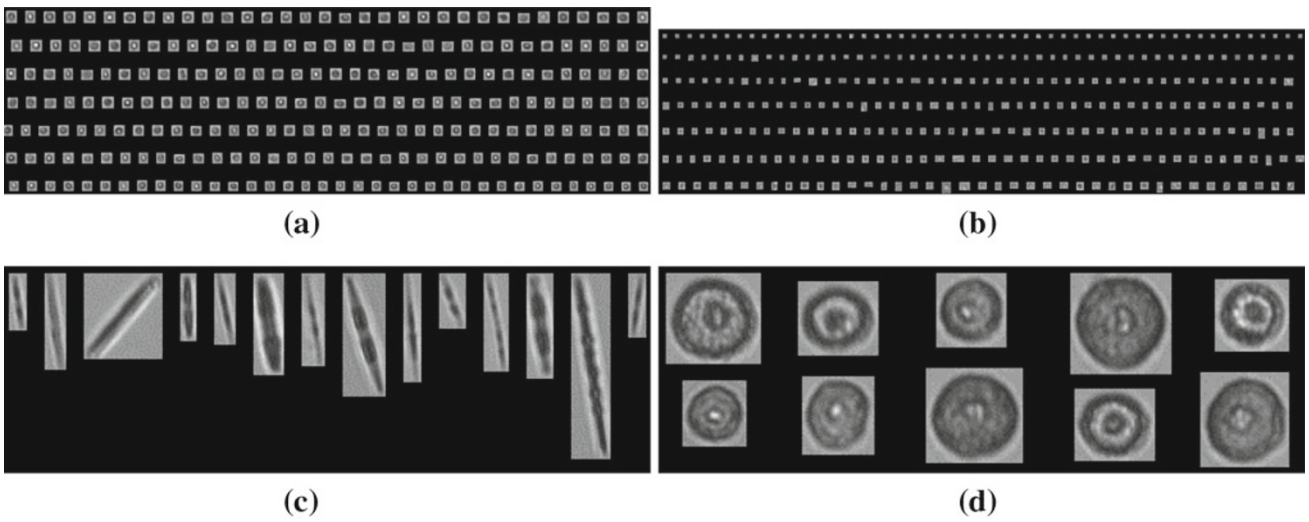

Fig. 4 Examples of the four classes of microalgae acquired by FlowCAM on the first dataset. This dataset were classified on four different classes: **a** flagellates, **c** pennate diatoms, **d** mesopores, and **b** others. This *figure* shows some important characteristics of this data as the unbalance and the reduced information about each microalgae

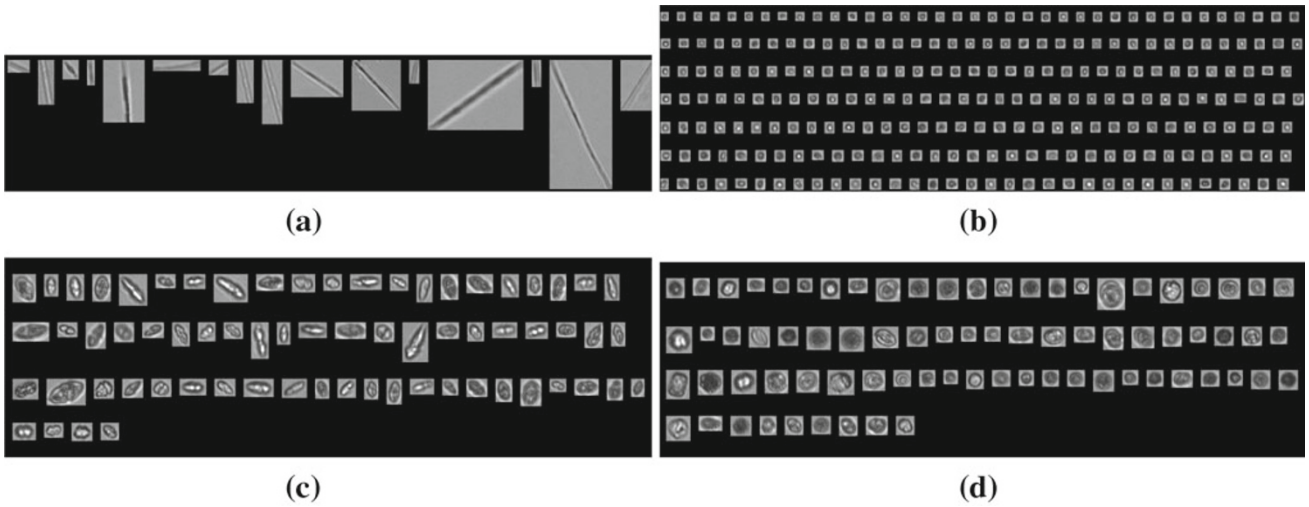

Fig. 5 Examples of the four classes of microalgae acquired by FlowCAM on the second dataset. This dataset were classified on four different classes: **a** pennate diatoms, **b** flagellates, **c** gymnodinium and **d** prorocentrales. As in the previous figure, it shows some important characteristics of this data as the unbalance and the reduced information about each microalgae

samples. Thus, using both metrics, the F1-score was evaluated. It is called maxF1 when obtained using *macro-average* and minF1 when obtained using *micro-average*. In the case of multiple classes, the minF1 has the same value as the metric known as accuracy (Ac), which is defined by Eq. 13. Thus, this work uses these two metrics: accuracy, or minF1, and maxF1.

$$\mathrm{Ac} = \frac{\sum_k Tp_k}{N}, \tag{13}$$

where N is the total number of samples on the data base and $\sum_k$ is the sum for all classes.

Some results were obtained in order to validate the approach using these two datasets completely classified. The first dataset is composed by 1,526 microalgae divided into four classes, as previously described, each one with 1,003 (Flagellates), 500 (others), 14 (pennate diatoms) and 9 samples (mesopores). The second dataset is composed by 923. It is also divided in four classes, as previously described, each one with 112 (Pennate Diatoms), 669 (Flagellates), 65 (gymnodinium) and 77 samples (prorocentrales).

From these datasets, smaller classified bases were randomly generated, with approximately 1, 3, 5, 10, 20 and 50 % of the original dataset, where each class should have at least one sample. In order to obtain quantitative results, for each percentage were generated ten different instances. Forty-eight samples were actively selected and classified.

4.1 Evaluation of the active learning

Firstly, the active learning capabilities were evaluated using three different metrics: *Least Confidence Sampling*, *Margin Sampling* e *Entropy*, when compared with a random selection. The results obtained in the first dataset is shown in Fig. 6. It shows the results for 1, 10 and 50 % of initial supervision using both evaluation metrics: Accuracy and maxF1.

Figure 6a and b show the results for 1 % of initial semi-supervision, in which the random selection presents a small accuracy and maxF1 raises with the addition of new samples. On the other hand, the other metrics had a significant improvement, especially on accuracy, which means a better classification independently of the classes.

On Fig. 6c and d, the results for 10 % of initial supervision are shown. It can be noted that the accuracy starts at a higher value than 1 % of semi-supervision and increases smoother for all metrics and the random selection. For the results obtained with 50 %, the variance is even smaller, as shown in Fig. 6e and f. Moreover, in this case, the active learning presents a small improvement for accuracy and maxF1.

The random selection can be seen as a semi-supervised addition of samples, thus, it can be noted that the active learning represents a significant gain, especially when there is little initial information.

Considering the second dataset, the results are similar to the ones obtained with the first dataset. One important difference between the datasets is the mean of accuracy and maxF1. The second dataset has different classes of microalgae, and the intraclass variability is larger than the first dataset. Thus, it is hardier to classify.

Figure 7 shows the results for 1, 3 and 5 % of initial supervision. It is interesting to see in Fig. 7a and b that the random selection of samples to classify in the active learning presents better results than the statistical techniques. This improvement happens after five samples and remains until the end of

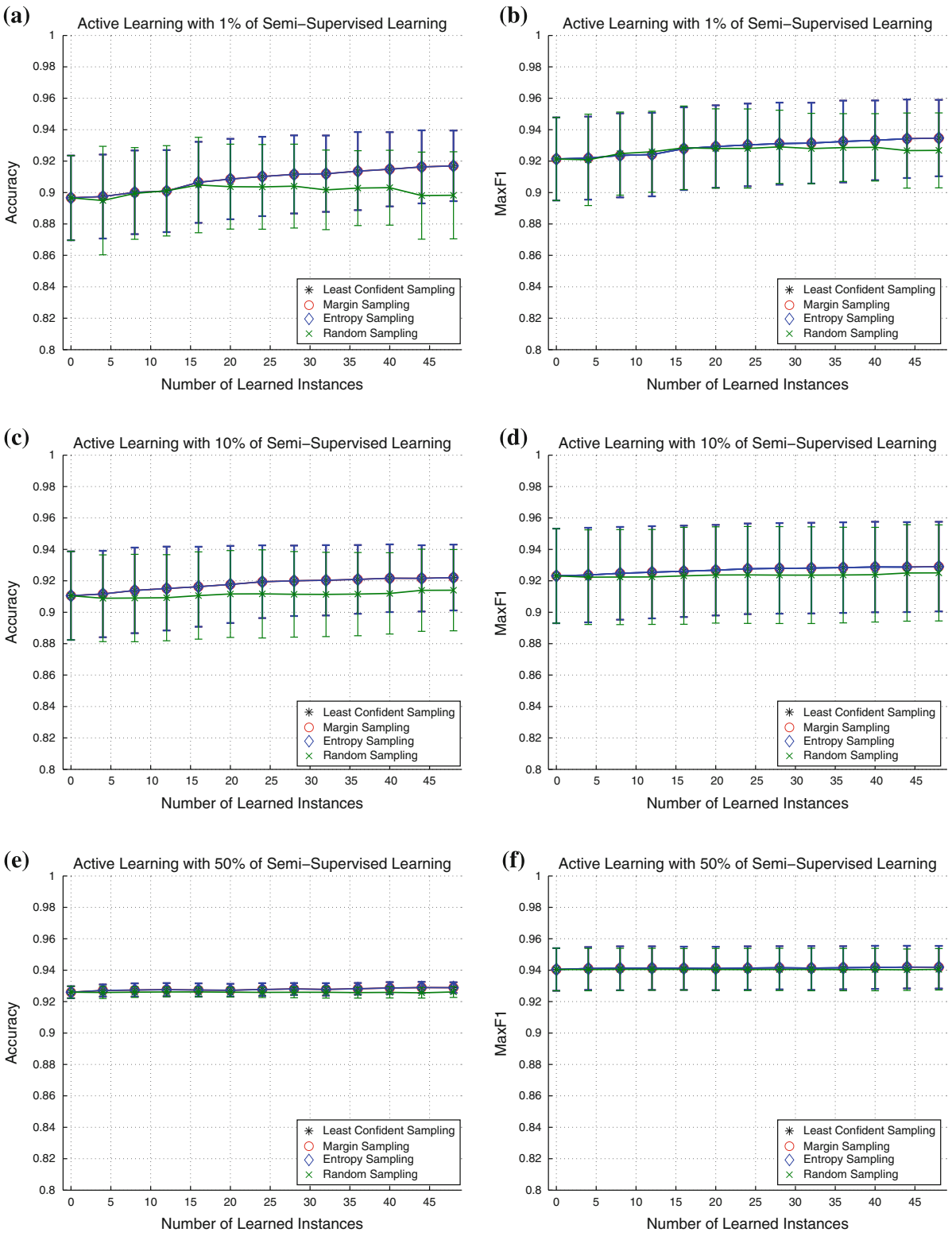

Fig. 6 Comparative of active learning metrics against a random selection using the *first dataset*, with results showing mean and standard deviation for the datasets with ten different bases. *The vertical axis* represents the accuracy and *the horizontal axis* represents the number of active samples informed to the system. **a** Accuracy for 1 % of initial semi-supervision, **b** MaxF1 for 1 % of initial semi-supervision, **c** accuracy for 10 % of initial semi-supervision, **d** MaxF1 for 10 % of initial semi-supervision, **e** accuracy for 50 % of initial semi-supervision, **f** MaxF1 for 50 % of initial semi-supervision (color figure online)

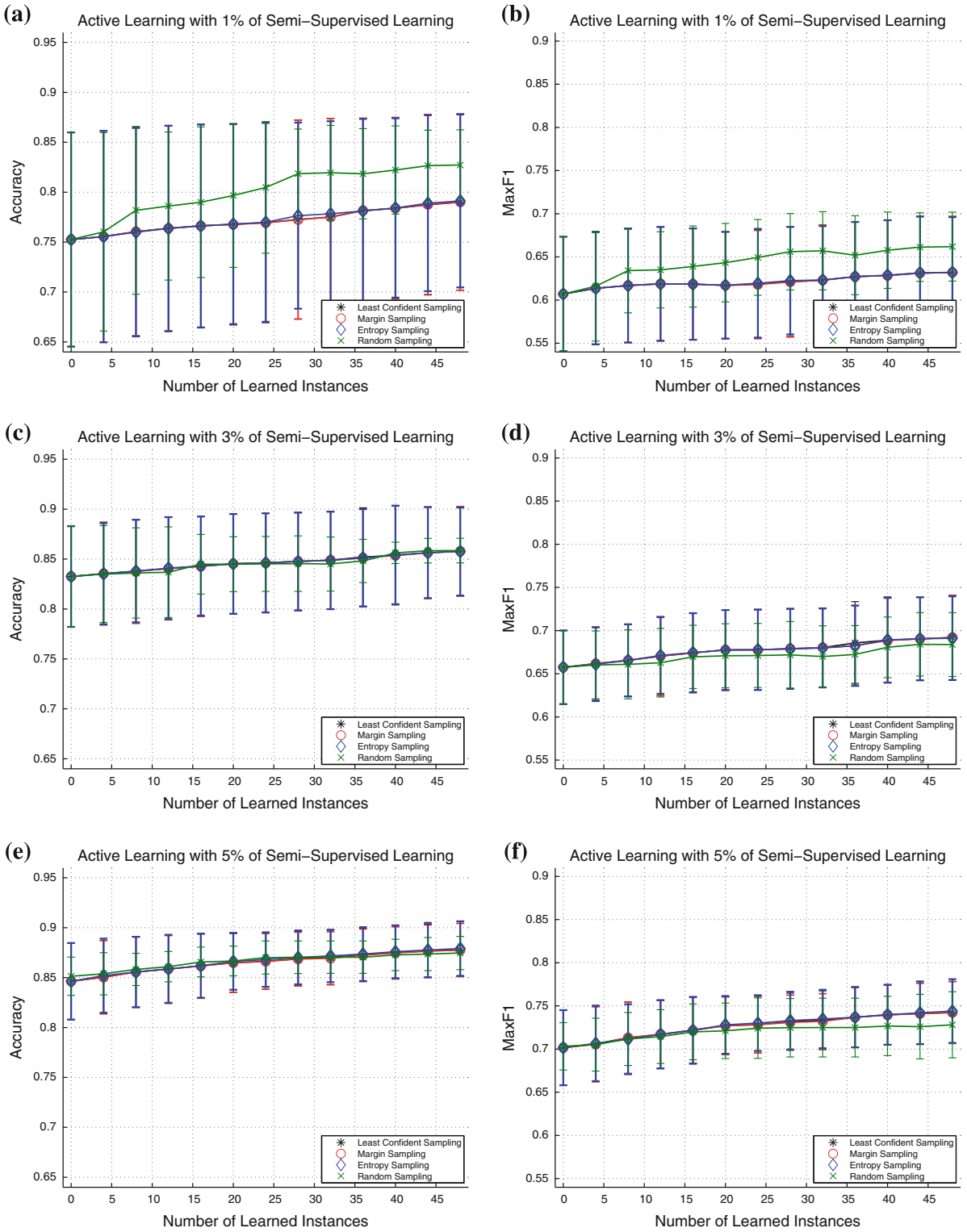

Fig. 7 Comparison of active learning metrics against a random selection using the *second dataset*, with results showing the average and standard deviation for the datasets with ten different instances. *The vertical axis* represents the accuracy and *the horizontal axis* represents the number of active samples informed to the system. **a** Accuracy for 1 % of initial semi-supervision, **b** MaxF1 for 1 % of initial semi-supervision, **c** accuracy for 3 % of initial semi-supervision, **d** MaxF1 for 3 % of initial semi-supervision, **e** accuracy for 5 % of initial semi-supervision, **f** MaxF1 for 5 % of initial semi-supervision (color figure online)

the active learning. The main reason for these results is due the large intraclass variance in this dataset. Thus, the system is not able to classify with a small number of supervised samples. In this case, the statistical selection falls into "local minima". In this case, the random selection chooses samples that improve the results, while the statistical methods choose samples that obtain a small improvement in relation to the random one.

In Fig. 7c and d, the results obtained by all selection methods are similar, with the maxF1 metrics of the random selection being worse than the others are. Considering 5 % of initial supervision the statistical methods are better than random selection, as shown in Fig. 7e and f. This results remain to the 10, 20 and 50 %. The entropy based sampling obtains a small advantage to the other metrics in all cases.

Figure 8 shows the results for accuracy of each of the ten generated bases from both datasets, considering a semi-supervision of 1, 3 and 10 %. The vertical axis represents the accuracy and the horizontal axis represents the number of active samples informed to the system. The accuracy results for the semi-supervised learning can be seen at number zero of the horizontal axis. As expected, the results shows that with a small semi-supervision, the accuracy is ruled by the chosen samples, and as the number of active samples increases, the variance decreases. The first two columns in this figure show the results using entropy for both datasets, and the last column using random selection for the second dataset.

On Fig. 8a and b, it is possible to see the difference in accuracy obtained by the proposed methodology for both dataset using 1 % of supervision. The second dataset is hardier to classify, thus the results show bases with approximately 65 % in accuracy. In these two figures is possible to see an interesting characteristics of 1 % supervision, some samples are capable to improve the accuracy in more than 5 %, as the base in black in Fig. 8a and in blue in Fig. 8b.

This phenomenon also happens, in Fig. 8c, in a large scale, where the random selection is used. It mainly occurs in bases where the initial accuracy is smaller. Thus, this base is composed by unrepresentative instances. Therefore, a representative sample can improved the capability of the system to classify correctly the unclassified data. It explains the results obtained in the Fig. 7a and b.

Figure 8d and e show the results for each base, considering 3 % of supervision. In this case, the first dataset shows a better accuracy value than the second dataset. The characteristics of the results are similar, with almost all bases with a small increasing in the accuracy with the active learning. Moreover, both results presents a base with small initial accuracy. This base, as previously explained, is sensible to random selection that generated some steps in the accuracy, as shown in Fig. 8f. The other bases are less sensible to the random selection, where the increasing in the accuracy is almost zero for the random selection, by the other side; the entropy based sampling is able to select good samples that increases the accuracy.

On Fig. 8d, there are two extreme cases. The first one, on cyan, 92 % of accuracy is obtained with a small supervision, while, on the second one, on red, only 86 % of accuracy is obtained. It can be noted that all instances had an improvement when new samples are actively selected. This is clearer on the red instance that goes from 86 % to almost 90 %. This fact also occurs in both datasets.

The results obtained using 10 % of supervision with entropy selection is shown in Fig. 8g and h. As seen in the previous results, the first dataset present a better accuracy than second dataset. The new actively selected samples improve in a small way the classification. This is due to a small capacity of generalization for this big group of samples, which means that new samples adds little information. It is also possible to see in Fig. 8i, the random selection presents a very small improvement in the accuracy. In this case, there are only a small number of informative samples to be selected, and the random method has a small chance of selecting them. Although these facts, the entropy- based sampling select informative samples. It is shown by the increasing in the accuracy of almost all bases, in Fig. 8g and h.

In order to evaluate whether the obtained performance was adequate, the results were compared with the support vector machine (SVM) [3] algorithm. This algorithm is considered the state-of-art on supervised learning and classification. The libSVM implementation [2] was used with a radial base kernel function, which presented better results. All other parameters were kept to its default.

Figure 9 shows the results including the active learning as a supervision addition. In black are shown the results obtained using only semi-supervised learning. The results after using the active learning are shown in red, which makes the supervision percentage to be raised. The blue line links the data used on the initialization of the active learning after forty-eight instances, in percentage.

Figure 9a and b shows semi-supervised learning, active learning using entropy and SVM learning results to accuracy and MaxF1 metrics for the first dataset. It can be noted that SVM has a small accuracy improvement with the increase of supervision, although it has better results than the semi-supervised algorithm alone. Only for 50 % of semi-supervision the presented approach obtains a better accuracy, while the active learning has similar results to the ones obtained by SVM.

On the other hand, the proposed approach presents superior results of maxF1. It is due to the unbalance between the classes. The SVM had excellent classification for the flagellates class, which has 1,003 samples, but did not have any sample classified for mesopores class, which has only 9 samples. This has a reflect on the accuracy and maxF1 metrics, as the accuracy metric only cares for the number of samples

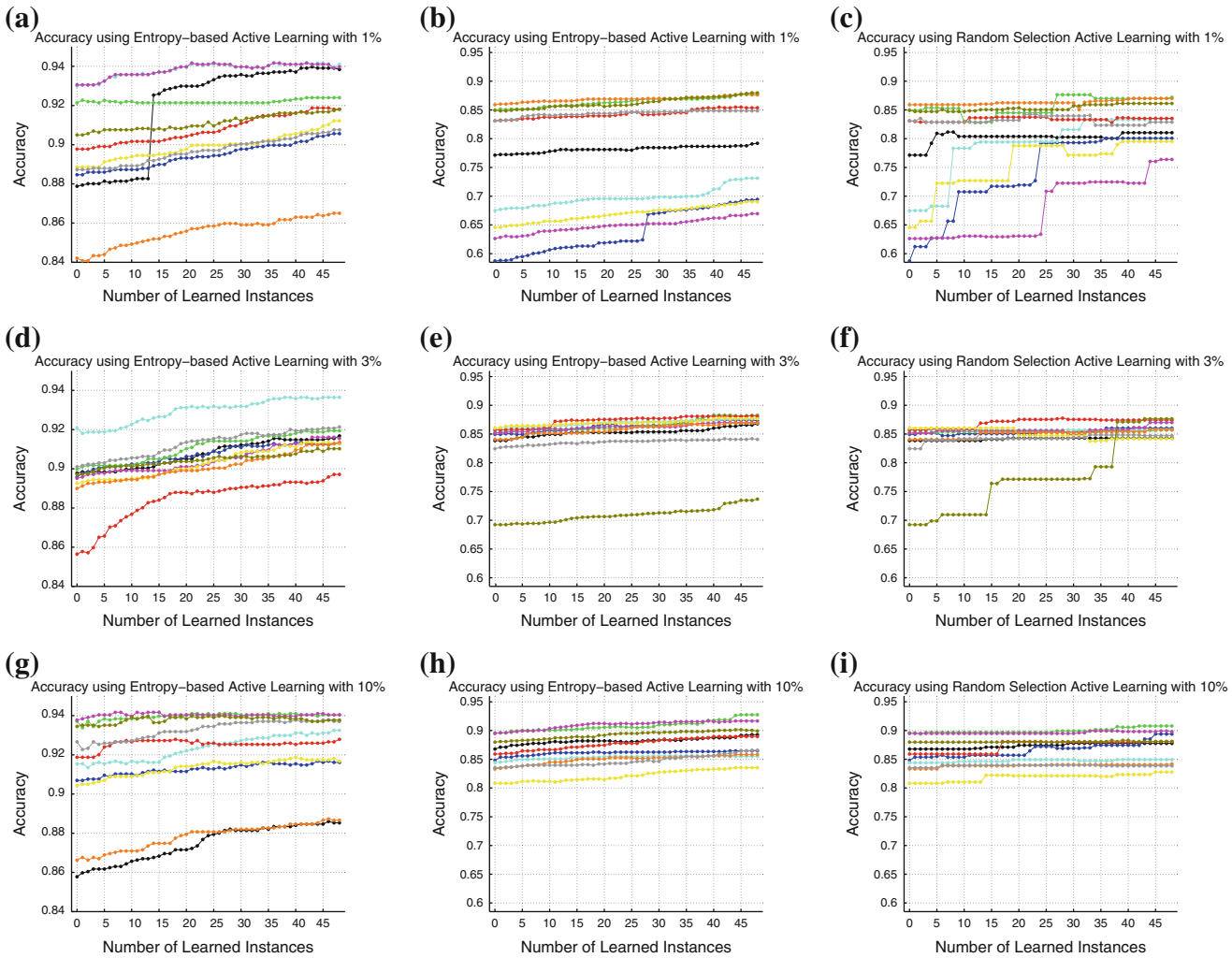

Fig. 8 Evaluation of the different instances of the semi-supervised data. In order to obtain statistical results, we generated ten different instances for each supervision percentage. The accuracy for all ten instances is shown using the proposed method with entropy, in the first two columns, and random selection, in the last column. The visualization is improved using different colors. *the vertical axis* represents the accuracy and *the horizontal axis* represents the number of active samples informed to the system. It is important to call attention to the vertical axis, where the intervals are different between the results from the first and the second datasets. **a** Results for 1 % of semi-supervision in the *first dataset*. **b** Results for 1 % of semi-supervision in the *second dataset*. **c** Results for 1 % of semi-supervision in the *second dataset*. **d** Results for 3 % of semi-supervision in the *first dataset*. **e** Results for 3 % of semi-supervision in the *second dataset*. **f** Results for 3 % of semi-supervision in the *second dataset*. **g** Results for 10 in the *first dataset*. **h** Results for 10 % of semi-supervision in the *second dataset*. **i** Results for 10 % of semi-supervision in the second dataset (color figure online)

that were correct classified, while the maxF1 cares for the number of samples classified for each class. In addition, it is interest of researchers to classify samples on all classes, especially the ones with small number of microalgae.

It is possible to notice that the gain obtained by the active learning is reduced as the semi-supervision increases. This effect happens with both metrics, accuracy and maxF1.

For the second dataset, the results are similar, but there are small differences, as shown in Fig. 9c and d. Due the large intraclass variance, the SVM obtains a better accuracy only until 5 % of supervision, after it, the semi-supervised algorithm obtains a better results. The maxF1 metric shows the main problem of the SVM results. The method has difficult to correct classify unbalanced datasets. But, it is a natural characteristics in this kind of dataset.

The accuracy obtained in the second dataset is smaller than the first one. However, the accuracy for bases with 50 % of supervision in the second dataset is greater than the first dataset, where after the active learning the accuracy is 95 %. Differently of the first dataset, the maxF1 continues increasing after active learning, even after 10 % of initial supervision, as shown Fig. 9d. It can be seen be the inclination of the blue line at 50 % between semi-supervised learning and the active learning.

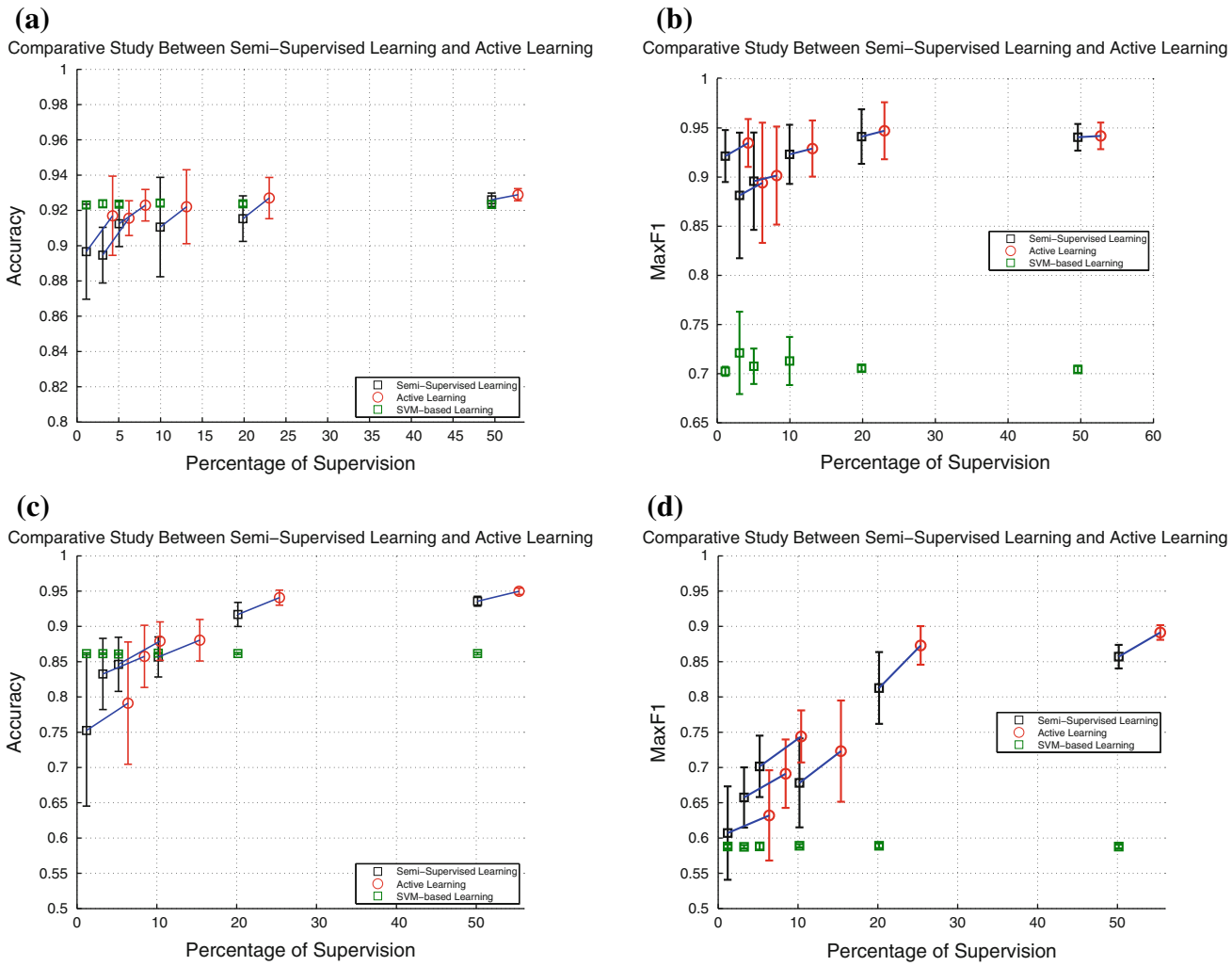

Fig. 9 Comparative of semi-supervised learning, *in black color*, against active learning using entropy, *in red color*. It is important that the semi-supervised learning be used as initial step to the active learning. Therefore, the semi-supervised results is s linked to the active learning by a *blue line*. The results obtained using SVM method are trained from the same supervised data used in semi-supervised approach, in *green color*. The mean and standard deviation are estimated and illustrated in the figure. Two different metrics are evaluated: accuracy and maxF1. **a** Accuracy comparative in the *first dataset*. **b** MaxF1 comparative in the *first dataset*. **c** Accuracy comparative in the *second dataset*. **d** MaxF1 comparative in the *second dataset* (color figure online)

5 Conclusion

This work proposed an approach for the classification of microalgae using a combination of semi-supervised and active learning algorithms. At the proposed approach, the semi-supervised classification is done using Gaussian mixture models together with the *expectation-maximization* algorithm. This classification is improved by the use of an active learning.

Two metrics, accuracy and maxF1, were used to validate the proposed approach, which presented favorable results for both metrics, achieving around 92 % of accuracy. The approach was compared with a state of the art algorithm of supervised learning, SVM, presenting similar results of accuracy and much better results of MaxF1. In this work, we presented three information evaluation metrics for the active learning, which had similar results with a small advantage to the entropy based sampling. The results show that the use of active learning improves the accuracy and the maxF1 with few samples.

The results obtained are relevant because, according to Culverhouse et al. [6], the hit rate achieved by humans remains between 67 and 83 %.

As a future direction, we intend to verify other methods capable of dealing with a larger number of classes and data, in order to generate a database of classified microalgae. Another improvement is to develop an adaptive model that automatically determines the number of classes. Finally, the features

obtained by the FlowCAM are limited, and as shown in this work, have problems concerning segmentation of microalgae. Thus, we will study image processing approaches to improve the segmentation and increase the amount of relevant features of the samples.

References

1. Blaschko MB, Holness G, Mattar MA, Lisin D, Utgoff PE, Hanson AR, Schultz H, Riseman EM, Sieracki ME, Balch WM, Tupper B (2005) Automatic in situ identification of plankton. In: IEEE workshops on application of computer vision (WACV), Breckenridge, Co, USA, pp 79–86
2. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(3):1–27
3. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
4. Costello MJ, Coll M, Danovaro R, Halpin P, Ojaveer H, Miloslavich P (2010) A census of marine biodiversity knowledge, resources, and future challenges. PLoS One 5(8):e12,110
5. Cullen JJ, Franks PJS, Karl DM, Longhurst A (2002) Physical influences on marine ecosystem dynamics. In: The sea, vol 12, chap 8, pp 297–336
6. Culverhouse P, Williams R, Reguera B, Herry V, Gonzlez-Gil S (2003) Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. Mar Ecol Prog Ser 247:17–25
7. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39(1):1–38
8. Drews-Jr P, Núñez P, Rocha R, Campos M, Dias J (2010) Novelty detection and 3 D shape retrieval using superquadrics and multi-scale sampling for autonomous mobile robot. In: Proceedings of the IEEE international conference on robotics and automation-ICRA, Anchorage, Alaska, USA, pp 3635–3640
9. Drews-Jr P, Colares RG, Machado P, de Faria M, Detoni A, Tavano V (2012) Aprendizado ativo e semi-supervisionado na classificacao de microalgas (in portuguese). In: IX Encontro Nacional de Inteligência Artificial-ENIA, Curitiba, Brazil
10. Drews-Jr P, Silva S, Marcolino L, Núñez P (2013) Fast and adaptive 3D change detection algorithm for autonomous robots based on Gaussian mixture models. In: Proceedings of the IEEE international conference on robotics and automation-ICRA, Karlsruhe, Germany, pp 4670–4675
11. Fluid Imaging Technologies Inc (2011) FlowCAM manual, 3rd edn. 65 Forest Falls Drive, Yarmouth, Maine, USA
12. Friedman A, Steinberg D, Pizarro O, Williams SB (2011) Active learning using a variational Dirichlet process model for pre-clustering and classification of underwater stereo imagery. In: IEEE/RSJ international conference on intelligent robots and system-IROS, IEEE, pp 1533–1539
13. Hamilton P, Proulx M, Earle C (2001) Enumerating phytoplankton with an upright compound microscope using a modified settling chamber. Hydrobiologia 444(1):171–175
14. Hu Q, Davis C (2006) Accurate automatic quantification of taxa-specific plankton abundance using dual classification with correction. Mar Ecol Prog Ser 306:51–61
15. Jakobsen H, Carstensen J (2011) FlowCAM: sizing cells and understanding the impact of size distributions on biovolume of planktonic community structure. Aquat Microb Ecol 65:75–87
16. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8): 1226–1238. doi:10.1109/TPAMI.2005.159
17. van Rijsbergen CJ (1979) Information retrieval. Butterworth-Heinemann, Glasgow
18. Settles B (2009) Active learning literature survey. Technical Report 1648, Computer Sciences. University of Wisconsin-Madison
19. Sosik HM, Olson RJ (2007) Automated taxonomic classification of phytoplankton sampled with imaging in-flow cytometry. Limnol Oceanogr Methods 5:204–216
20. Veloso A, Meira W Jr, Cristo M, Gonçalves M, Zaki M (2006) Multi-evidence, multi-criteria, lazy associative document classification. In: ACM international conference on Information and knowledge management, pp 218–227
21. Xu L, Jiang T, Xie J, Zheng S (2010) Red tide algae classification using SVM-SNP and semi-supervised FCM. In: 2nd International conference on education technology and computer-ICETC, pp 389–392
22. Zhu X, Goldberg AB (2009) Introduction to semi-supervised learning. Synth Lect Artif Intell Mach Learn 3(1):1–130